NetKarma Provenance Retrieval and Visualization Plug-in
For Cytoscape
User Manual
V1.1.0 August 21, 2011

NetKarma

DATA TO INSIGHT CENTER
INDIANA UNIVERSITY
Pervasive Technology Institute

Copyright 2011 The Trustees of Indiana University

This document contains instructions for using the NetKarma provenance retrieval and visualization plug-in version 1.1.0, which provides core capability to retrieve provenance information from a NetKarma provenance system and visualize the returned graph. NetKarma provenance retrieval and visualization plug-in is licensed under Apache License, Version 2.0 (the "License") (http://www.apache.org/licenses/LICENSE-2.0). The code is copyrighted and copyright owned by The Trustees of Indiana University. NetKarma provenance retrieval and visualization plug-in is a product of the Data to Insight Center at Indiana University. See http://pti.iu.edu/d2i/provenance for more information.

Contents

1. Introduction 4
2. Software Dependencies 4
   2.1 Service dependencies 4
   2.2 Installation dependencies 5
3. Deploying Plug-in 5
   3.1 Deploying under Mac/Linux/Unix OS 5
   3.2 Deploying under Windows OS 6
4. Configuring the Plug-in Properties 6
   4.1 Updating the configuration files to connect NetKarma provenance retrieval and visualization plug-ins to NetKarma service 6
   4.2 The configuration file for the NetKarma visualization plug-in 7
5. Using the NetKarma Provenance Retrieval Plug-in 9
6. Using the NetKarma Visualization Plug-in 11
   6.1 Load the NetKarma provenance XML file 11
   6.2 Applying layout algorithms to the NetKarma provenance graph 13
   6.3 Navigating the NetKarma provenance graph 14
   6.4 Play movie 15
   6.5 Save as XML 16
   6.6 Get data provenance history 17
   6.7 X axis sort 18
   6.8 Creating an abstract view 19
      6.8.1 Clustering neighbor nodes 19
      6.8.2 Navigate between graph and sub-graphs 22
      6.8.3 Compress Process/Artifact 23

1. Introduction

We have developed two plug-ins to Cytoscape to visualize and navigate provenance information contained in the NetKarma provenance system.

The NetKarma provenance retrieval plug-in is used to retrieve provenance graphs from a remote NetKarma server to your local machine. In the Cytoscape tool, this plug-in is run by clicking on the blue "Karma" icon in the Cytoscape tool bar after the plug-in is installed. The second plug-in is the NetKarma visualization plug-in which allows you to visualize and manipulate graphs downloaded using the retrieval plug-in. The NetKarma visualization plug-in is displayed as an orange "Geni" icon in the Cytoscape toolbar.

Visualization of provenance data is useful for manipulating very large provenance graphs, for displaying different views, and for interactivity. This can help a user to navigate their experiment information with a mental map of what is going on in the experiment, to compare different experiment runs quantitatively, and to do model selection with an effective collaboration between the user and the discovery system.

NetKarma is a standalone system that can be added to existing cyberinfrastructure for purposes of collection and representation of provenance data. The NetKarma query plug-in aims to provide a GUI component that queries provenance information of scientific experiments from the NetKarma provenance repository. The NetKarma server is accessible via either a webservice API or RabbitMQ enterprise bus, and our provenance retrieval plug-in supports both access methods.

Cytoscape (http://www.cytoscape.org/) is an open source software platform for complex network analysis and visualization. We use Cytoscape because of its support for detail and overlaying visualizations with additional annotations. We developed our visualization tool as a plug-in that can generate the provenance graph visualization directly from the NetKarma provenance information using an XML, and provide control of the navigation process.

2. Software Dependencies

2.1 Service dependencies

The NetKarma plug-ins retrieve provenance as graphs from the NetKarma provenance server. To use the Cytoscape plug-ins, you will need to either setup a NetKarma server or process a log file using an established NetKarma service such the Data to Insight Center's NetKarma service on the GRNOC server. To set up a NetKarma server, please refer to the NetKarma Provenance System user guide at http://pti.iu.edu/d2i/provenance_karma

If an instance of the NetKarma service (either hosted as a web service or as a standalone service using the RabbitMQ messaging bus) already exists, contact your system administrator for obtaining access.

2.2 Installation dependencies

The NetKarma provenance retrieval and visualization plug-in v1.0.0 has been tested with the following software packages on which it has a dependency. These packages will need to be installed before using the visualization plug-in:

1) Java Development Kit (JDK) v5 or v6
   http://java.sun.com
2) Cytoscape v2.8.1
   http://www.cytoscape.org/

3. Deploying Plug-in

Download the plug-in package as a zip file from:
http://pti.iu.edu/d2i/provenance_karma

3.1 Deploying under Mac/Linux/Unix OS

1) Unzip the plug-in package:

```
unzip NetKarmaVisPlugin.zip
```

2) Copy the jar files into the plug-ins directory under Cytoscape_v2.8.1

Install the NetKarma visualization plug-in:

```
cp NetKarmaVisPlugin/KarmaGraph.jar cytoscape-2.8.1/plugins/
```

Install the NetKarma provenance retrieval plug-in:

```
cp NetKarmaVisPlugin/KarmaRetrieval.jar cytoscape-2.8.1/plugins/
```

3) Copy the configuration files into the plug-ins directory under Cytoscape_v2.8.1

Create a new directory named “config” under cytoscape’s “plugins” directory:

```
mkdir cytoscape-2.8.1/plugins/config
```

Deploy the configuration file for the NetKarma visualization plug-in:

```
cp NetKarmaVisPlugin/config/pluginConfig.xml cytoscape-2.8.1/plugins/config
```

Deploy the configuration file for the NetKarma provenance retrieval plug-in:

```
cp NetKarmaVisPlugin/config/karmaQueryConfig.txt cytoscape-2.8.1/plugins/config
```

4) After you finish installing the two plug-ins, the “plugins” directory under Cytoscape_v2.8.1 will look like:

```
-Cytoscape_v2.8.1 HOME DIRECTORY
---plugins/
--------KarmaGraph.jar
--------KarmaRetrieval.jar
--------config/
------------karmaQueryConfig.txt
------------pluginConfig.xml
```

3.2 Deploying under Windows OS

1) Unzip the plugin package

```
unzip NetKarmaVisPlugin.zip
```

2) Copy the jar files into the plugins directory under Cytoscape_v2.8.1

Install the NetKarma visualization plug-in:

```
copy NetKarmaVisPlugin\KarmaGraph.jar "C:\Program files\cytoscape-2.8.1\plugins\"
```

Install the NetKarma provenance retrieval plug-in:

```
copy NetKarmaVisPlugin\KarmaRetrieval.jar "C:\Program files\cytoscape-2.8.1\plugins\"
```

3) Copy the configuration files into the plugins directory under Cytoscape_v2.8.1

Create a new directory named "config" under cytoscape's "plugins" directory:

```
mkdir "C:\Program files\cytoscape-2.8.1\plugins\config"
```

Deploy the configuration file for the NetKarma visualization plug-in:

```
copy NetKarmaVisPlugin\config\pluginConfig.xml "C:\Program files\cytoscape-2.8.1\plugins\config"
```

Deploy the configuration file for the Karma provenance retrieval plug-in:

```
copy NetKarmaVisPlugin\config\karmaQueryConfig.txt "C:\Program files\cytoscape-2.8.1\plugins\config"
```

4) After you finish installing the two plug-ins, the "plugins" directory under Cytoscape_v2.8.1 will look like:

```
-Cytoscape_v2.8.1 HOME DIRECTORY
---plugins/
--------KarmaGraph.jar
--------KarmaRetrieval.jar
--------config/
------------karmaQueryConfig.txt
------------pluginConfig.xml
```

4. Configuring the Plug-in Properties

4.1 Updating the configuration files to connect NetKarma provenance retrieval and visualization plug-ins to NetKarma service

The following describes how to configure the provenance retrieval and visualization plug-ins and connect them to a running NetKarma service.

You have two options (Axis2 webservice or RabbitMQ messaging system) to connect to a NetKarma server, and all of the configuration information is stored in the file "karmaQueryConfig.txt". This

configuration file is used and managed by both NetKarma provenance retrieval plug-in and the visualization plug-in. The configuration file will be loaded at the startup of either plug-in, and the configuration settings will be displayed in the configuration panel. You can choose which option you want to use in the configuration panel, and modify the default configurations.

You can directly modify the contents of the "karmaQueryConfig.txt" file or make configuration changes through the configuration panel in Cytoscape.
To set up the configuration for connecting via the Axis2 webservice:
There is only one property the needs to be set when using the Axis2 server:
axis2.serviceURL – enter the URL to the NetKarma v3.1.2 webservice

To set up the configuration for connecting via RabbitMQ messaging system:
There are several properties that must be set to connect using the RabbtiMQ server:
messaging.username – username of RabbitMQ
messaging.password – password of RabbitMQ
messaging.hostname – hostname or IP address
messaging.hostport – port

4.2 The configuration file for the NetKarma visualization plug-in

The"pluginConfig.xml" file defines which properties of process and artifact nodes in the provenance graph will be added as annotations to their respective types of nodes in the visualization. To identify node properties that should be added as annotations, an "attribute" element is added to the "process" or "artifact" section within the "importAttributes" element within the configuration file. Each attribute element contains "attributeName" and "importName" child elements. When visualizing the graph, node properties are checked to see if the URI for the property contains the substring specified as the attributeName of an attribute in the configuration file. If the substring is found in the URI, then the value of the child element within the value element is added as an annotation to the respective node in the visualization. The description for that annotation will be the value specified for the importName.

As an example, if a process node in your provenance graph has the following property element:

```
<v1:property uri="urn:tool:gush:peng-virtual-machine-12345-1296341651/Process_1/process-entity-workflowID">
    <v1:value>
        <workflowID>urn:tool:gush:peng-virtual-machine-12345-1296341651</workflowID>
    </v1:value>
 </v1:property>
```

And the following "attribute" element is added to the set of process attributes in your configuration file:

```
<attribute>
        <attributeName>workflowID</attributeName>
        <importName>import workflowID</importName>
</attribute>
```

Then the "workflowID" attributeName would match the property since "workflowID" is a substring of the property's URI. Since the value element within that property also contains a child element, the value

of that element (urn:tool:gush:peng-virtual-machine-12345-1296341651) would be added as an annotation for that node in the visualization, and the label for that annotation would be "importWorkflowID".

Properties for artifact nodes within the provenance graph differ from properties for process nodes in that multiple property values can be contained within a nested "block" element within a property. For artifact nodes, in addition to searching for the attributeName in the URI of a property, the visualization plug-in looks inside the block element of a property to compare the tags for child elements within the block to the attributeNames specified for artifact "attributes" in the configuration file.

As an example, if your provenance graph contains the following property element for an artifact node:

```
<v1:property uri="urn:tool:gush:peng-virtual-machine-15555-1292253699/Block_3/artifact-content">
        <v1:value>
          <content>
            <dataBlocks xmlns:ns="http://www.dataandsearch.org/karma/2010/08/">
              <block>
                <objectID>urn:output:stdout</objectID>
                <objectName>stdout</objectName>
                <objectValue>[Hello World]</objectValue>
                <objectType>BLOCK</objectType>
              </block>
            </dataBlocks>
          </content>
        </v1:value>
</v1:property>
```

And the "ImportAttributes" section of your configuration file contains the following artifact attribute definitions:

```
<attribute>
        <attributeName>objectID</attributeName>
        <importName>artifactID</importName>
</attribute>
<attribute>
        <attributeName>objectName</attributeName>
        <importName>artifactName</importName>
</attribute>
```

Then your provenance graph would contain two annotations for this node in the visualization; an annotation labeled "artifactID" with the value "urn:output:stdout" and an annotation labeled "artifactName" with the value "stdout".

The configuration file also contains a "visualAttributes" section that allows you to specify which attributes of the process and artifact nodes should used to label the nodes when visualized. The attributes to be used for process nodes and artifact nodes are listed separately within "processLabel" and

"artifactLabel" elements within the configuration file. Each of these elements contains one or more "attributeName" elements that specify a prioritized list of the attributes to be used to label the nodes.

As an example, if your configuration file contains the following "artifactLabel" element:

```
<artifactLabel>
        <attributeName>objectValue</attributeName>
        <attributeName>ID</attributeName>
</artifactLabel>
```

Then for each artifact node, if it has an objectValue attribute, the value of that attribute will be used as the label for that artifact in the visualization. If an artifact does not have an objectValue attribute, then the ID attribute will be used to label that node.

5. Using the NetKarma Provenance Retrieval Plug-in

1) Run Cytoscape and click on the Karma icon in the toolbar.

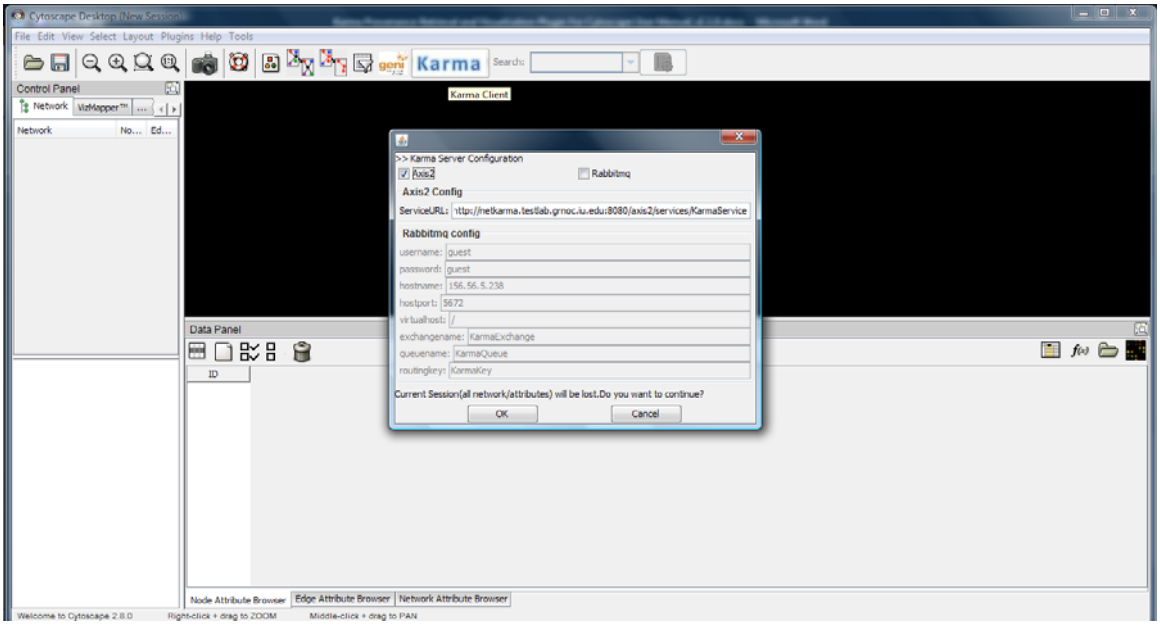

Figure 5.1 Click on NetKarma toolbar

If you have not already edited the configuration file (or wish to use different settings), choose the connection method you want to use and configure its connection parameters (Choosing "Axis2", the "ServiceURL" in the "Axis2 Config" panel needs to be configured; choosing "Rabbitmq", all parameters in the "Rabiitmq config" panel need to be configured).

2) When you are satisfied with your configuration settings, click the "OK" button and a new dialog window will appear which prompts you for the workflowID.

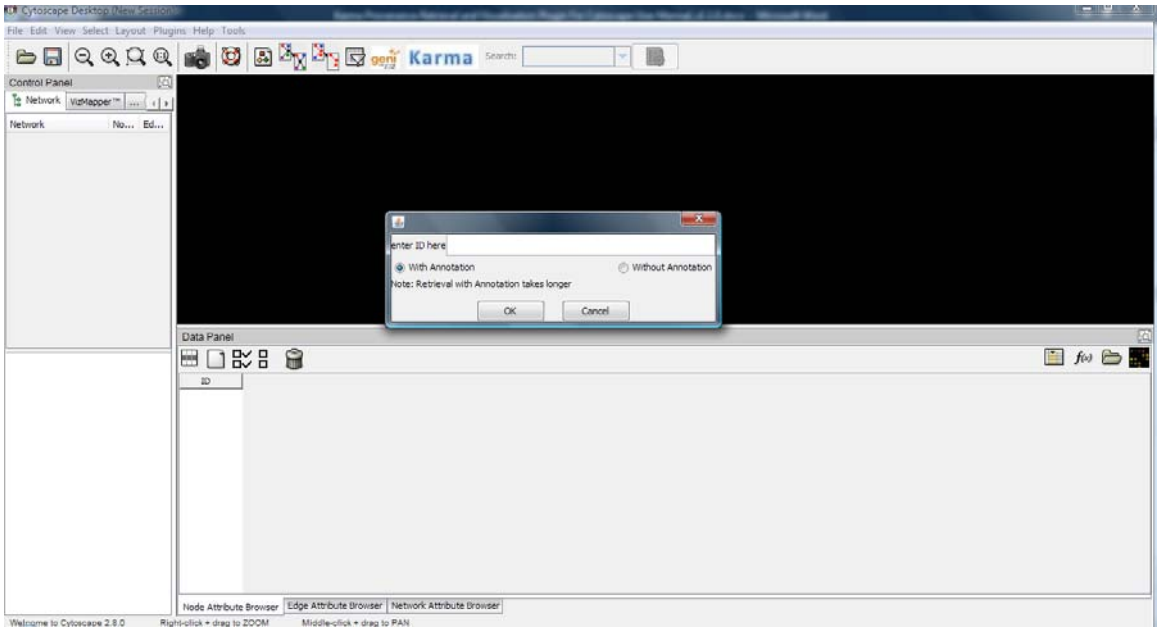

Figure 5.2 Enter the workflow ID

Enter the workflowID of the graph you want to retrieve, and select whether you want the annotations included in the graph or not. Press the OK button to continue.

3) After the plug-in has retrieved the NetKarma provenance XML, you will be prompted to choose the name and location to store that XMLfile.

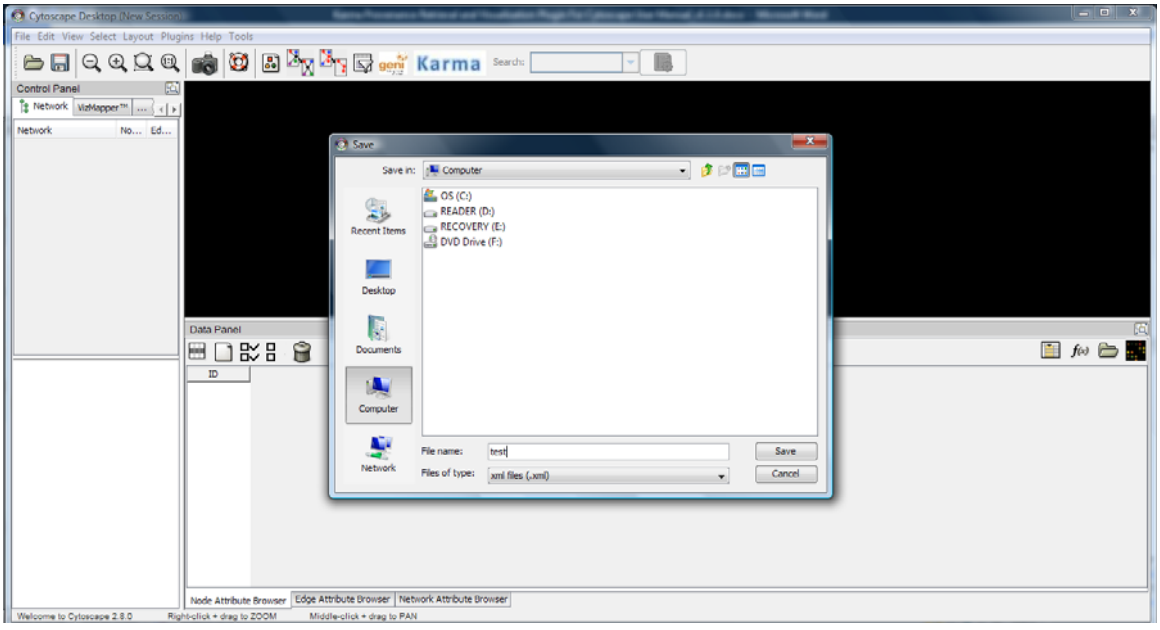

Figure 5.3 Save the OPM XML file

Note 1: The time required to retrieve the graph varies depending on the size and complexity of the XML file. Retrieving a graph without annotations can significantly reduce the time required. For large

provenance graphs we generally recommend retrieving without annotations since the NetKarma visualization plug-in will load the annotations “On Demand” while navigating the provenance graph.
Note 2: While retrieving the XML file, Cytoscape will appear to be blocked, but will continue to work right after the file is retrieved successfully. Please be patient while waiting for a graph to be downloaded.

6. Using the NetKarma Visualization Plug-in

6.1 Load the NetKarma provenance XML file

1) Click on the GENI toolbar

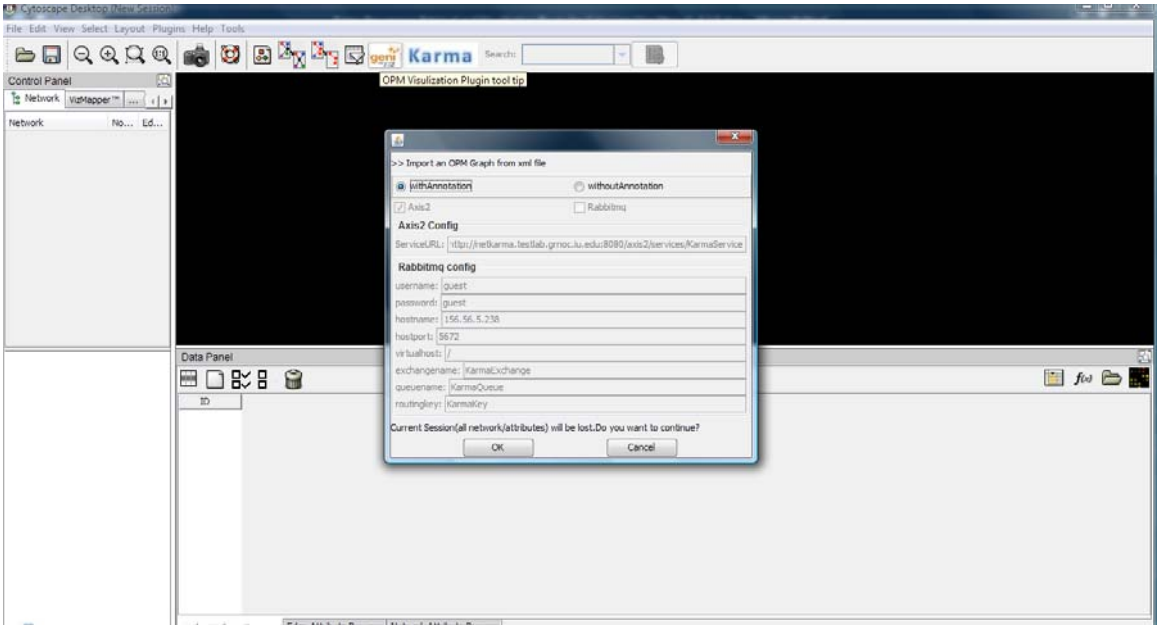

Figure 6.1.1 Click on the GENI icon in the toolbar

Select whether to import the provenance xml with annotations or without annotations. When you import provenance without annotations, the NetKarma visualization plug-in will set up a background connection to the NetKarma server, and load both annotations and the registry level information “On Demand” (when you select any node in the provenance graph during navigation, the NetKarma visualization will retrieve more information for it).

2) If you choose to import provenance without annotations, you also need to choose which connection method you want to use for the background connection and configure the connection parameters.

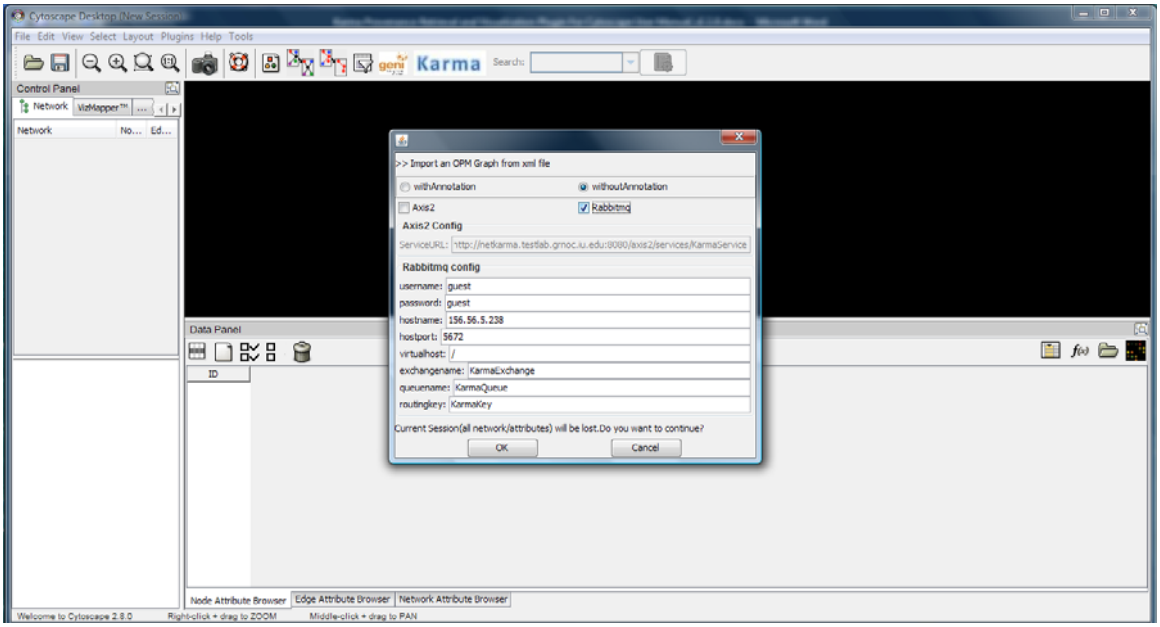

Figure 6.1.2 Configure the connection to Karma Server

3) Select the NetKarma provenance XML file

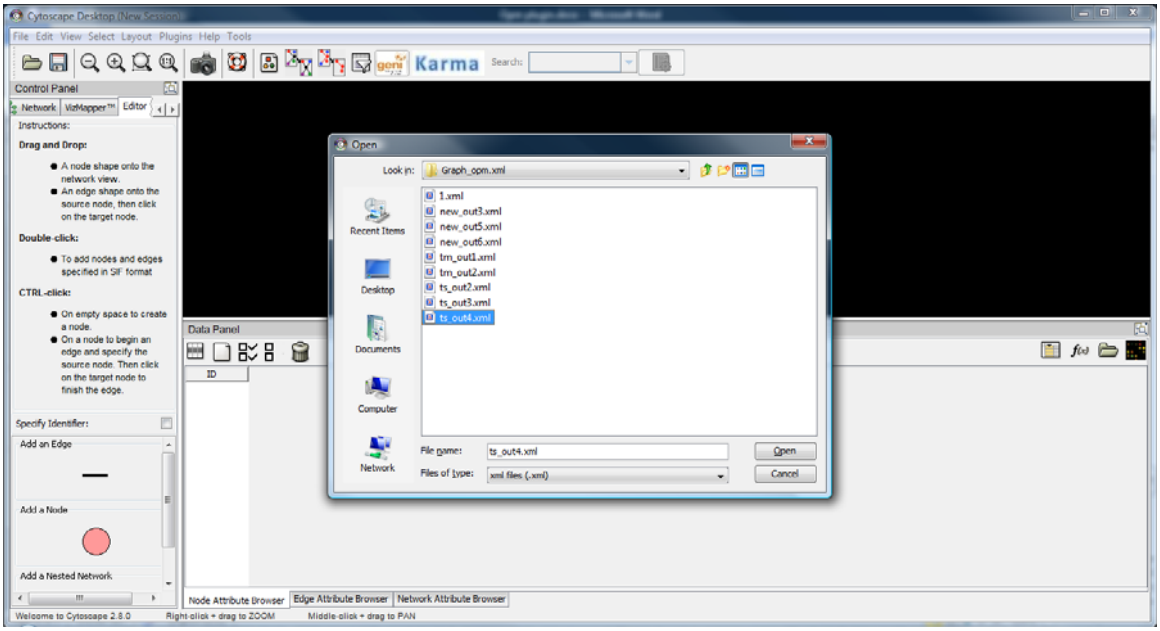

Figure 6.1.3 Select the NetKarma provenance XML file

4) Load the NetKarma provenance graph

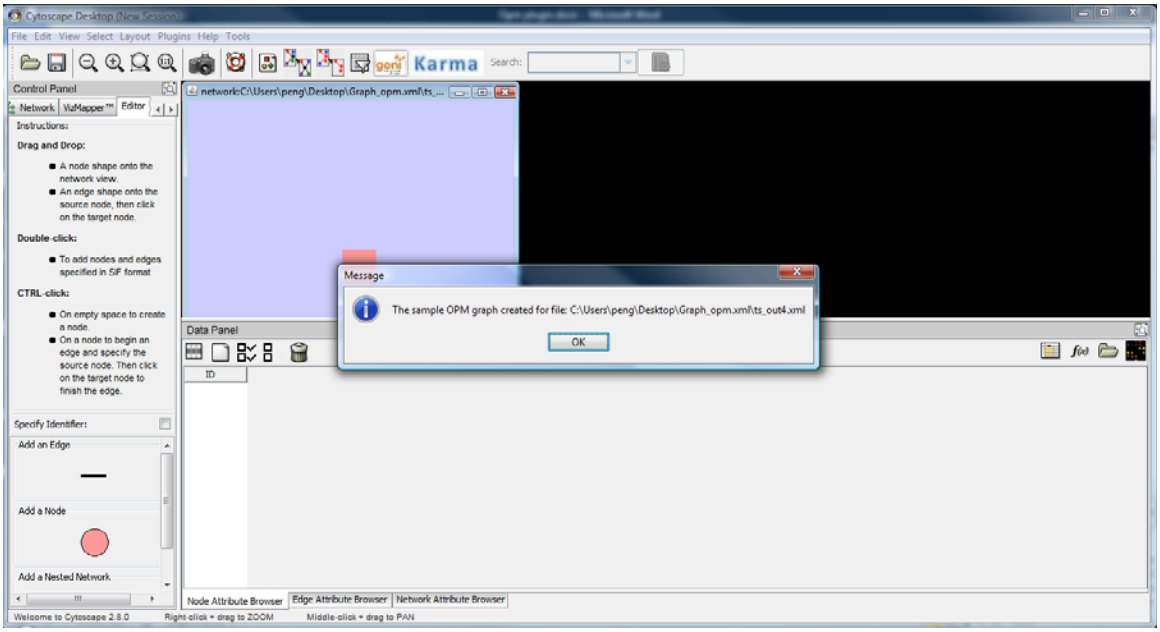

Figure 6.1.4 Load the NetKarma provenance graph

The initial graph will look like this (your graph may differ depending on your experiment's provenance):

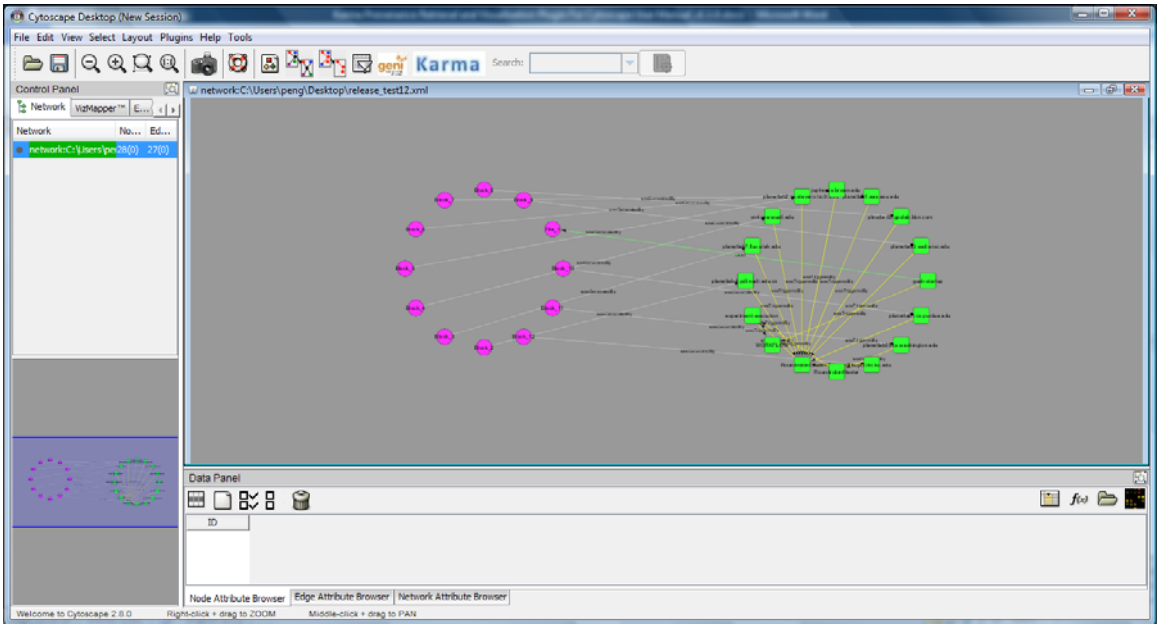

Figure 6.1.5 Initial graph

6.2 Applying layout algorithms to the NetKarma provenance graph

Different layouts for the provenance graph are available from the Cytoscape Layout/Cytoscape Layouts menu.

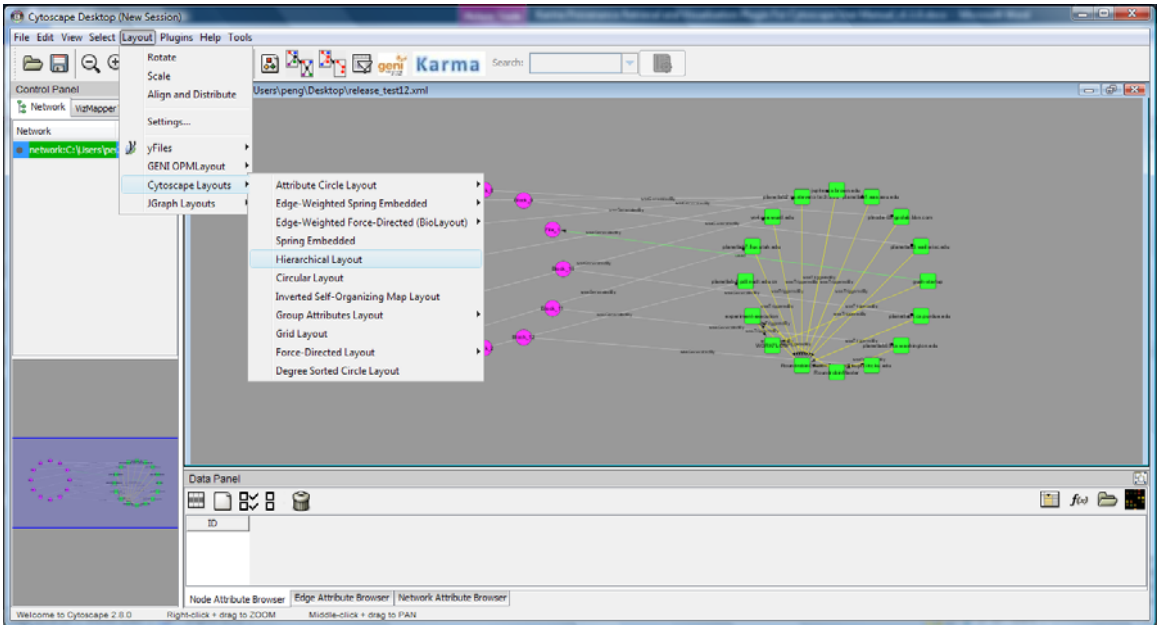

Figure 6.2.1 Do hierarchy layout

1) For example, if you select the hierarchy layout, the graph will appear as shown in Figure 6.2.2:

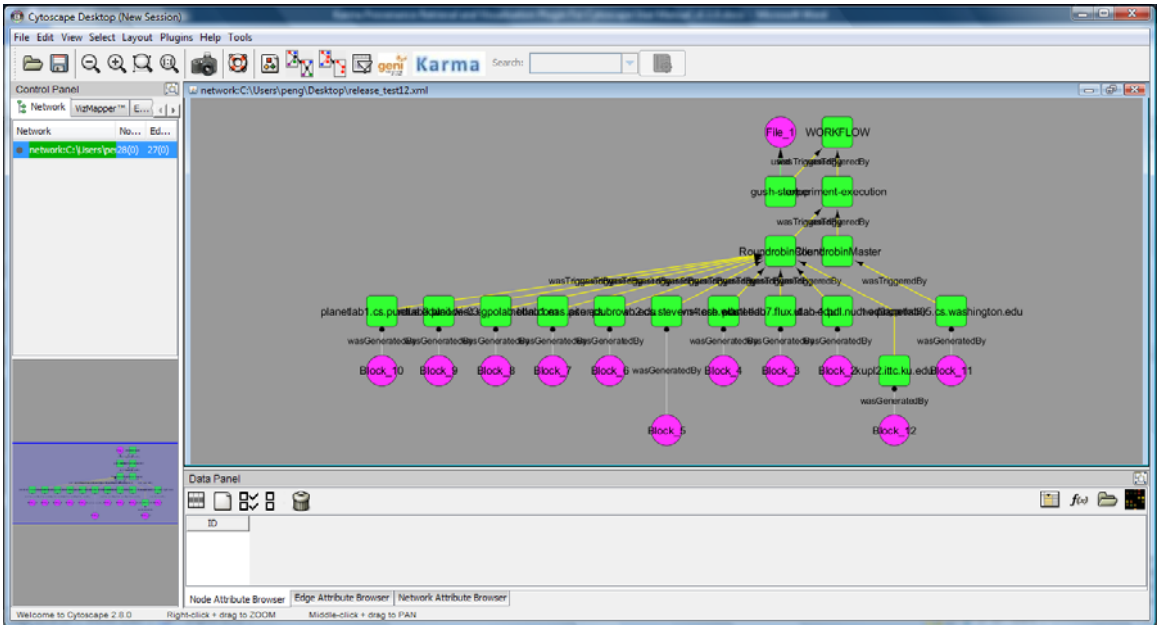

Figure 6.2.2 Hierarchical layout of graph

6.3 Navigating the NetKarma provenance graph

1) Double click on any node to see its attributes

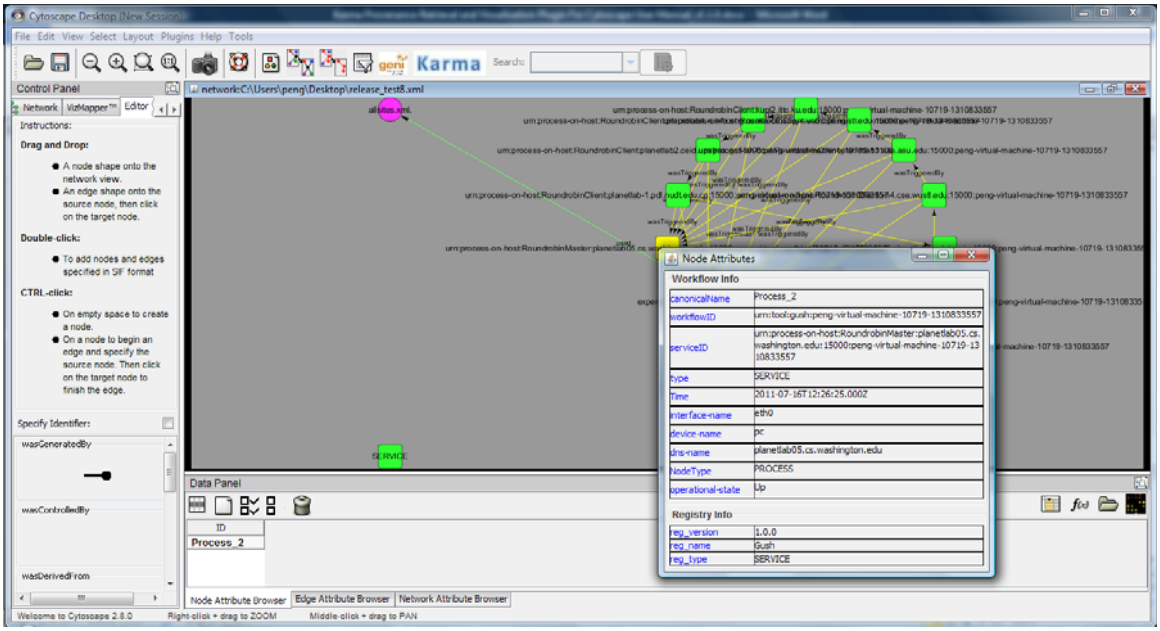

Figure 6.3.1 Double click on workflow node

2) Double click on any edge to see its attributes

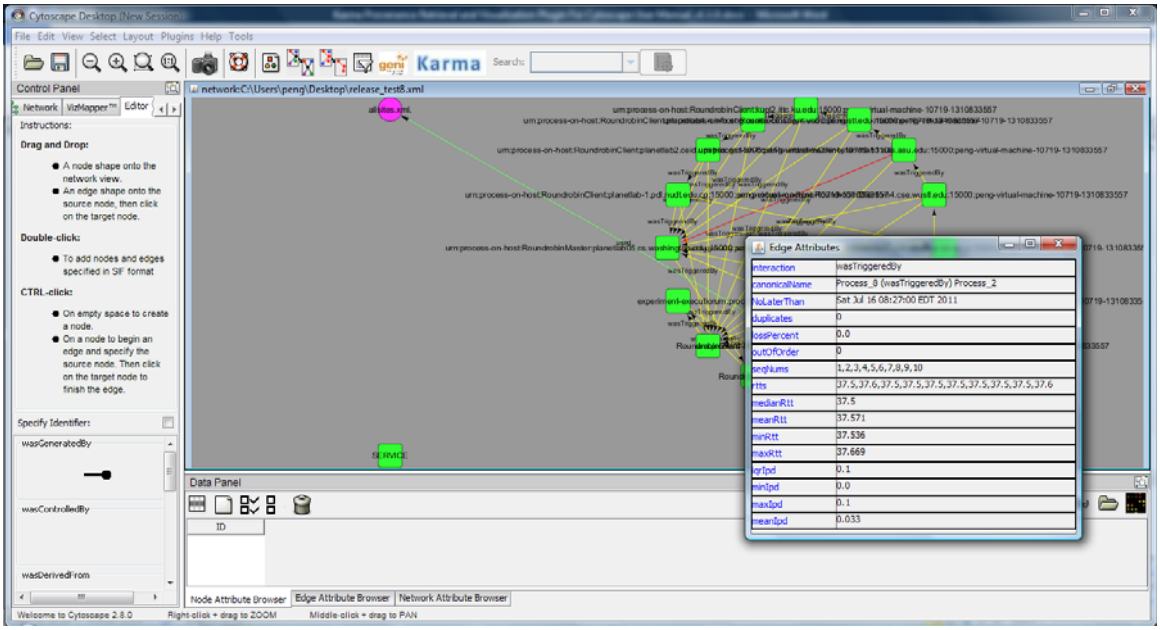

Figure 6.3.2 Double click on edge

6.4 Play movie

If you right click on any node, you can choose to play a movie displaying the provenance for the whole graph based on the order of the “timestep” or “time” attributes of the nodes.

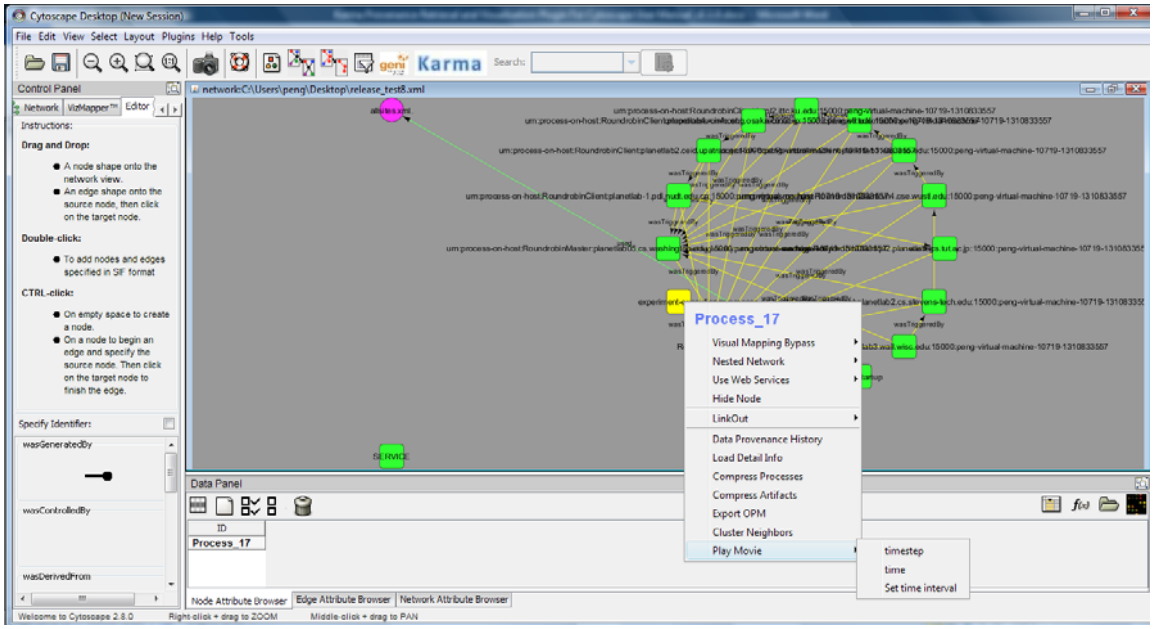

Figure 6.4.1 Display provenance as a movie

You can also configure the time interval (milliseconds) between the appearances of two nodes in the movie.

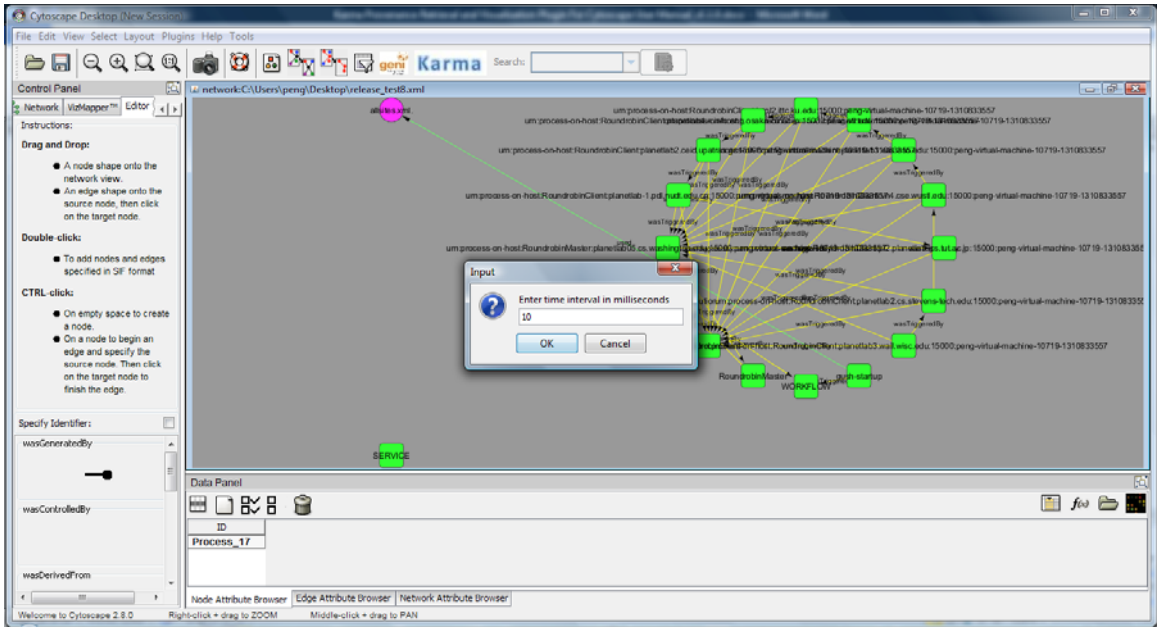

Figure 6.4.2 Configuring the time interval

6.5 Save as XML

To save the provenance graph displayed in Cytoscape as an XML file based on the Open Provenance Model (OPM), right click on any node and select the menu item “Export OPM”. You will be prompted for the location to save the XML file.

The graph’s structure will be based on the OPM v1.1 (Open Provenance Model, http://openprovenance.org/).

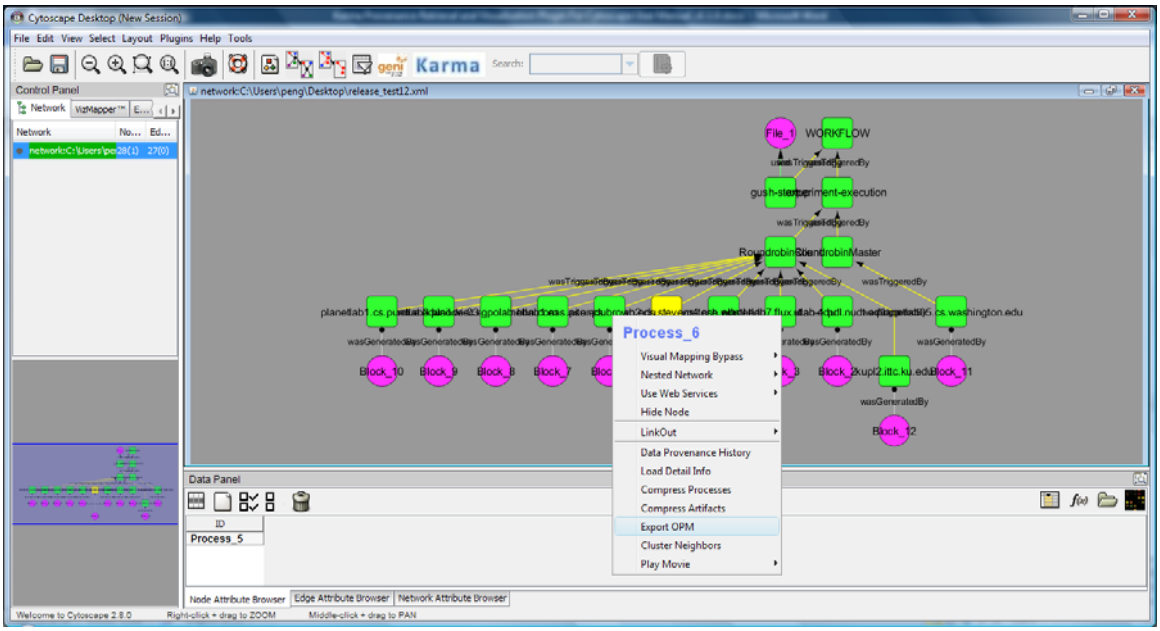

Figure 6.5.1 Click on “Export OPM”

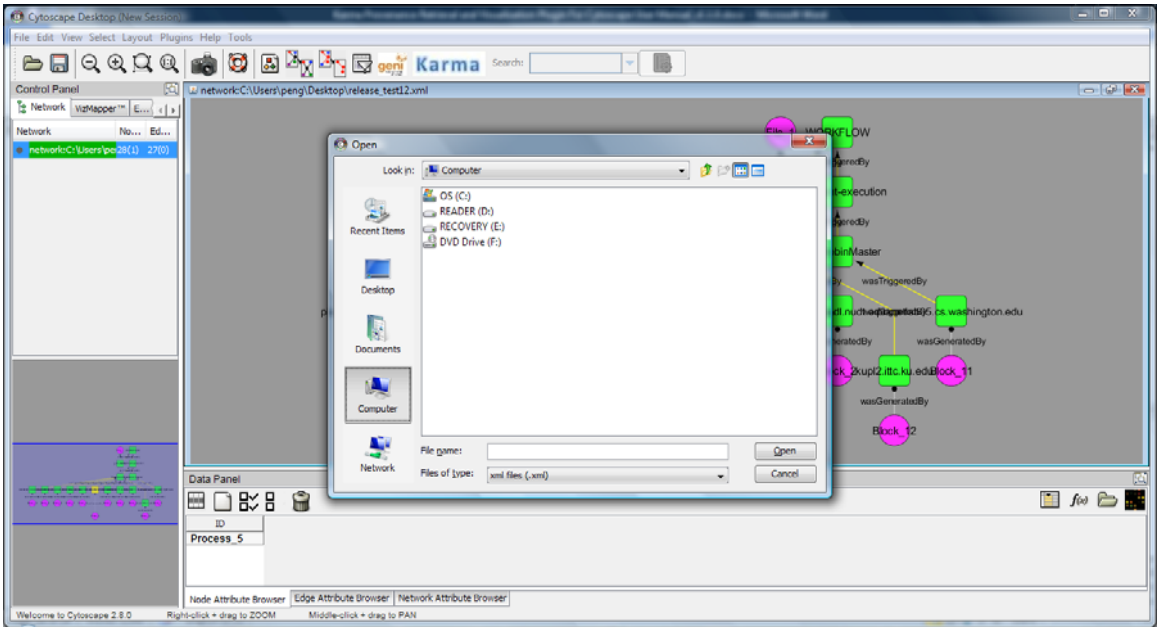

Figure 6.5.2 Choose the location to save the xml file

6.6 Get data provenance history

Right clicking on any artifact will display an option for “Data Provenance History”. Selecting that option will display a new graph that includes all of the nodes involved in the generation of the selected artifact (This method utilizes a background communication with the NetKarma server).

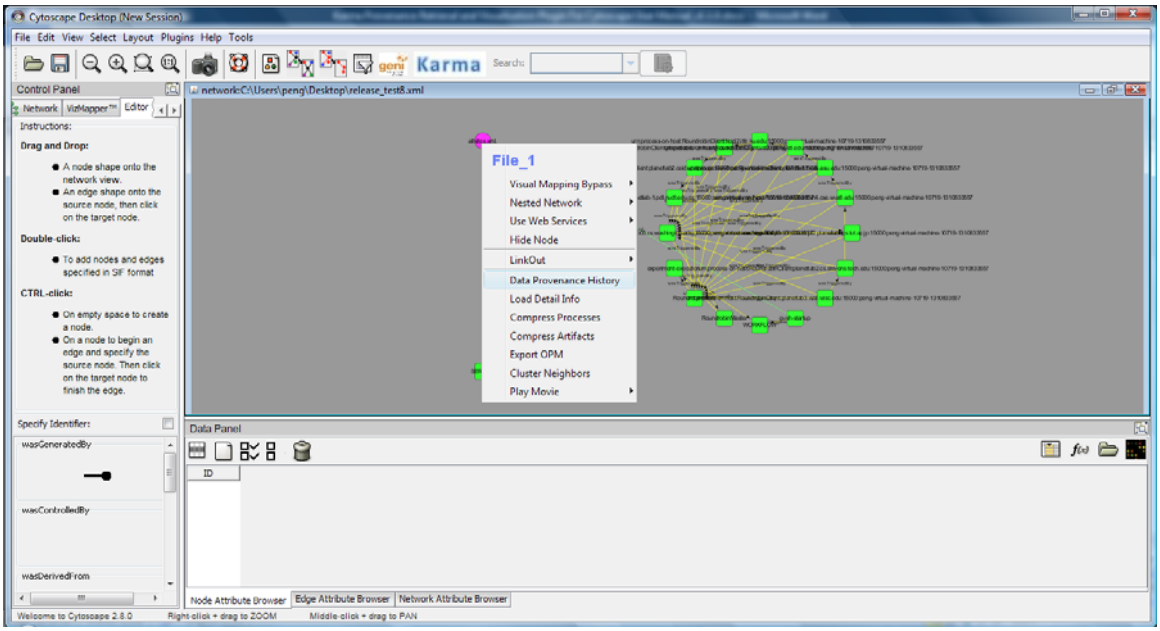

Figure 6.6.1 Get data provenance history

Note: This functionality works best when the NetKarma visualization plug-in is connected to NetKarma using the Axis2 webservice instead of the RabbitMQ.

6.7 X axis sort

You can display the nodes using a layered format based on the order of the “time” attribute by selecting the “OPM X sort” option from the Layout/GENI OPMLayout menu.

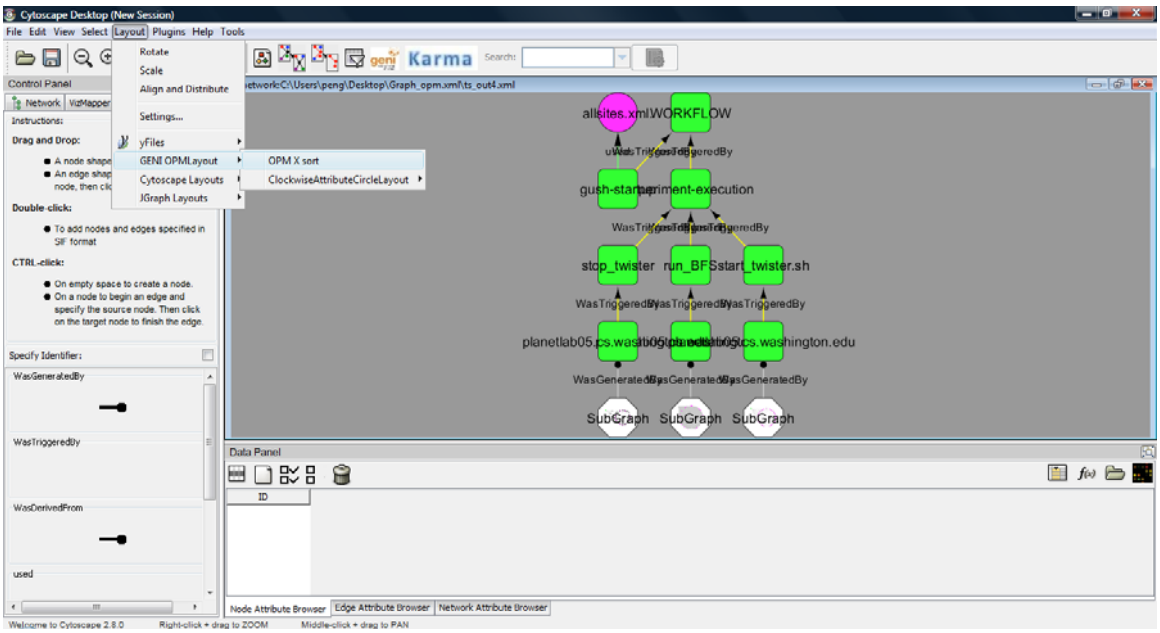

Figure 6.7.1 Select the X sort algorithm

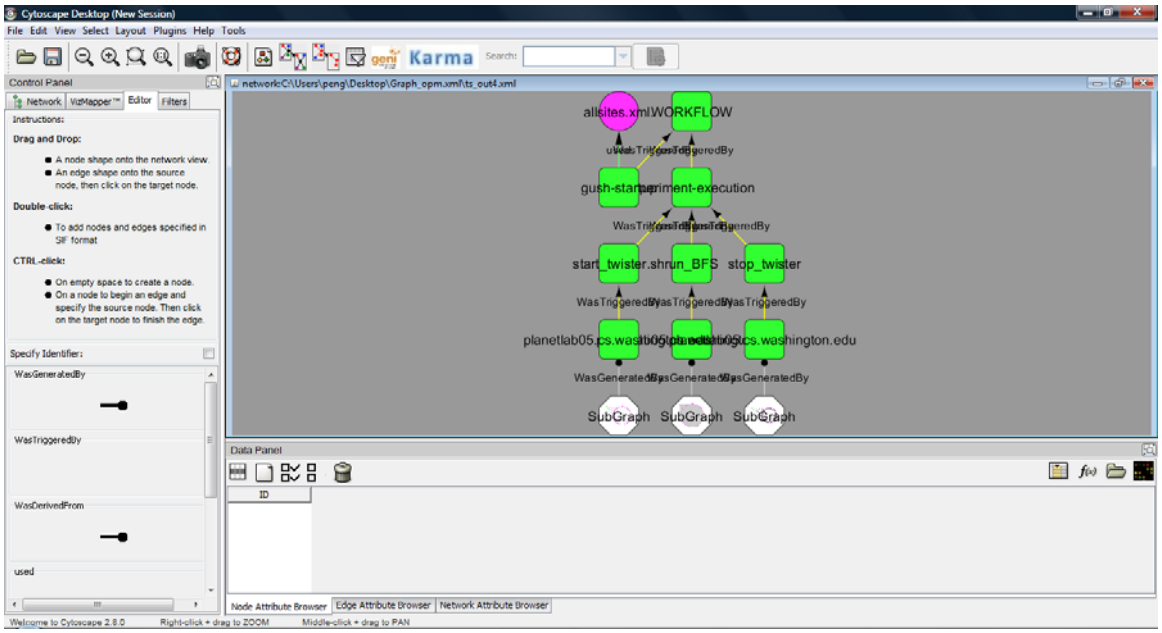

Figure 6.7.2 The hierarchy graph after sorting

6.8 Creating an abstract view

For some experiments, the complexity of the provenance relationships can result in very complicated graphs. This section introduces features of the NetKarma visualization plug-in that can be sued to abstract out some of this complexity to allow you to visualize specific aspects of the provenance graph.

6.8.1 Clustering neighbor nodes

To deal with graphs with a large number of "artifact" nodes, the NetKarma visualization plug-in supports extracting an abstract view by clustering neighboring nodes. As an example, Figure 6.8.1.1 shows a graph that has a larger number of "artifact" nodes generated by a small number of "process" nodes.

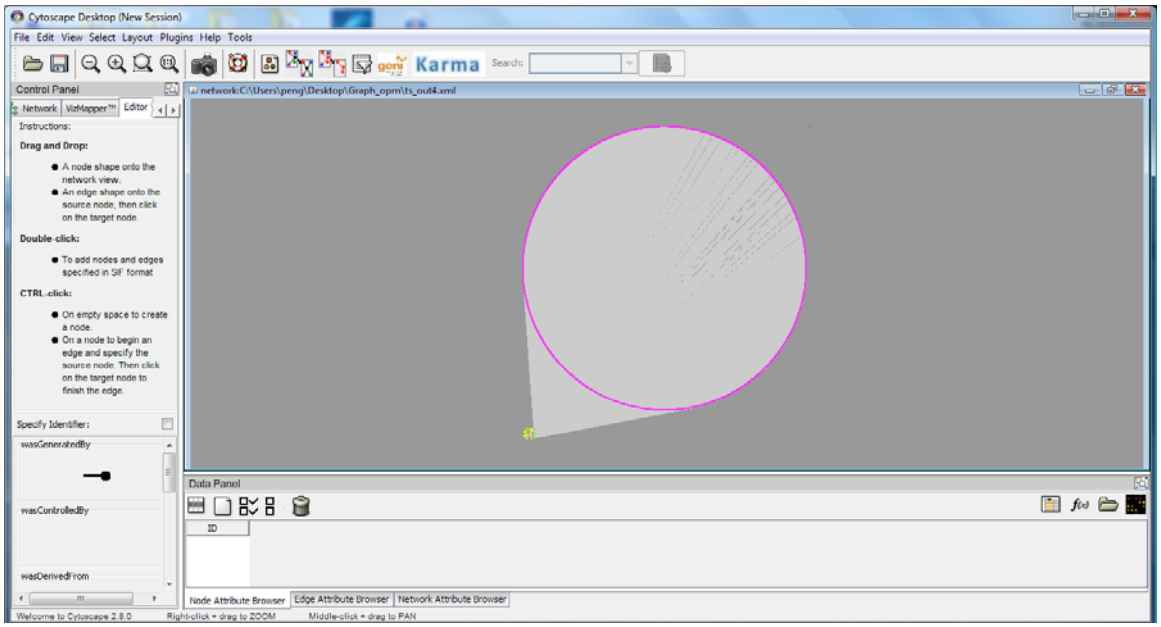
Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help Tools
Karma
Control Panel
Network VizMapper™ Editor
Instructions:
Drag and Drop:
A node shape onto the network view.
An edge shape onto the source node, then click on the target node.
Double-click:
To add nodes and edges specified in SIF format
CTRL-click:
On empty space to create a node.
On a node to begin an edge and specify the source node. Then click on the target node to finish the edge.
Specify Identifier:
wasGeneratedBy
wasControlledBy
wasDerivedFrom
Data Panel
ID
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.8.0
Right-click + drag to ZOOM
Middle-click + drag to PAN

Figure 6.8.1.1 A graph with large number of "artifact" nodes

You can create an abstract view for this type of graph using the following 5-step process:

1) Select the process nodes (on the lower left side of the graph in Figure 6.8.1.2):

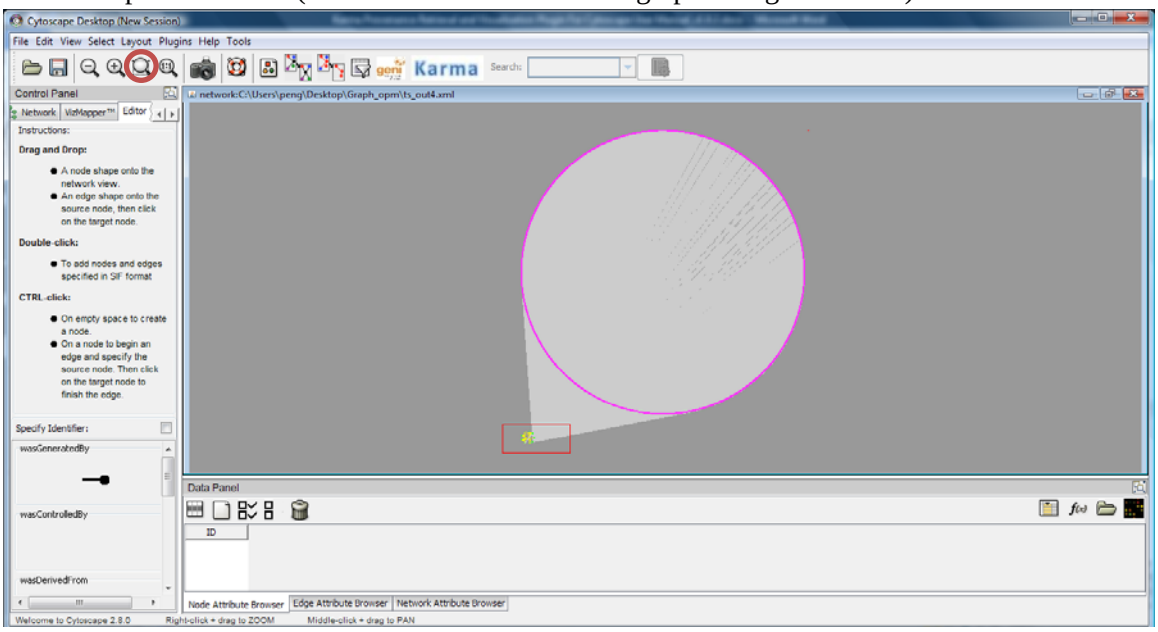
Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help Tools
Karma
Control Panel
Network VizMapper™ Editor
Instructions:
Drag and Drop:
A node shape onto the network view.
An edge shape onto the source node, then click on the target node.
Double-click:
To add nodes and edges specified in SIF format
CTRL-click:
On empty space to create a node.
On a node to begin an edge and specify the source node. Then click on the target node to finish the edge.
Specify Identifier:
wasGeneratedBy
wasControlledBy
wasDerivedFrom
Data Panel
ID
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.8.0
Right-click + drag to ZOOM
Middle-click + drag to PAN

Figure 6.8.1.2 Select processes

2) Click “zoom selected region” icon from the toolbar (highlighted magnifying glass icon the Figure 6.8.1.3):

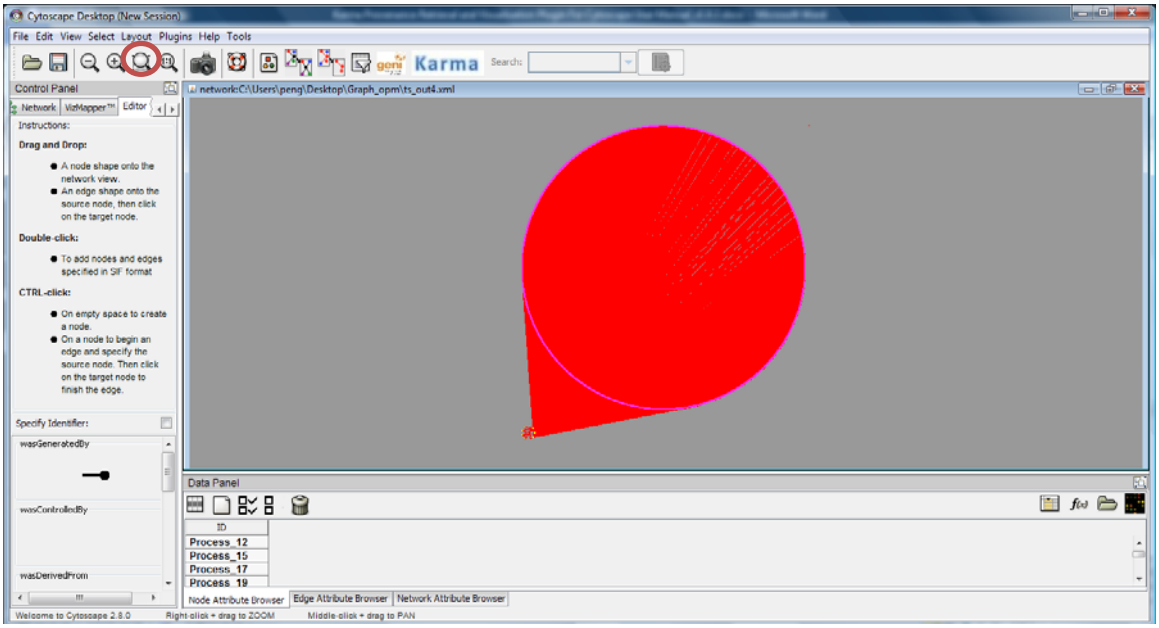

Figure 6.8.1.3 Zoom into process region

3) Right click on one of the nodes with a large out-degree (a significant number of edges):

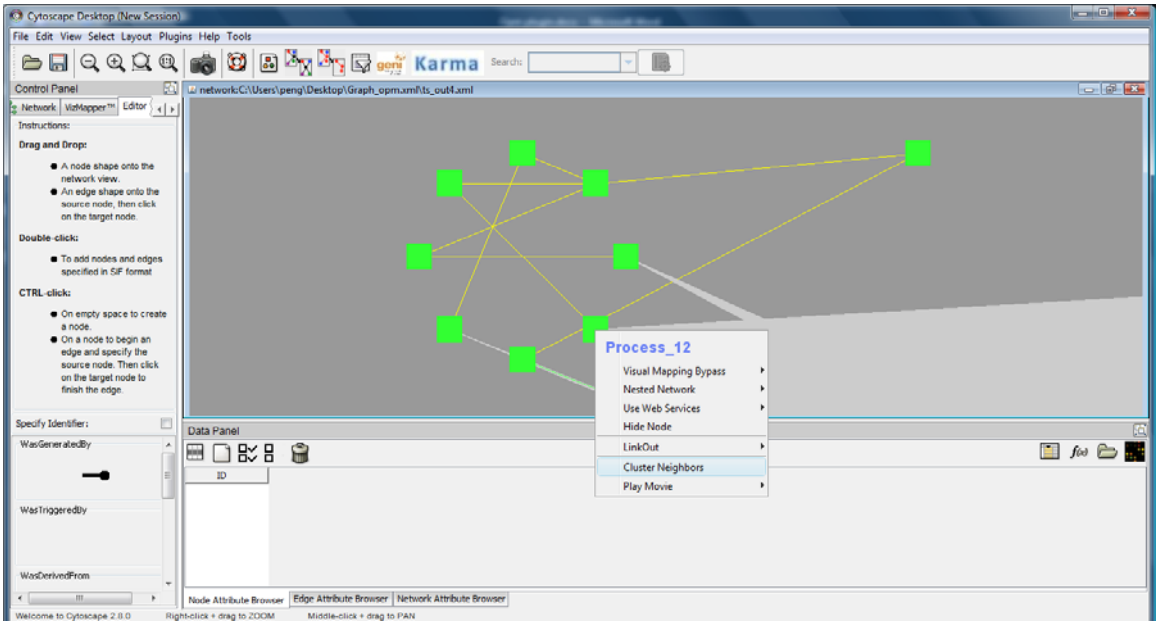

Figure 6.8.1.4 Right click on node

4) Clustering the neighbors of the node selected in step 3:

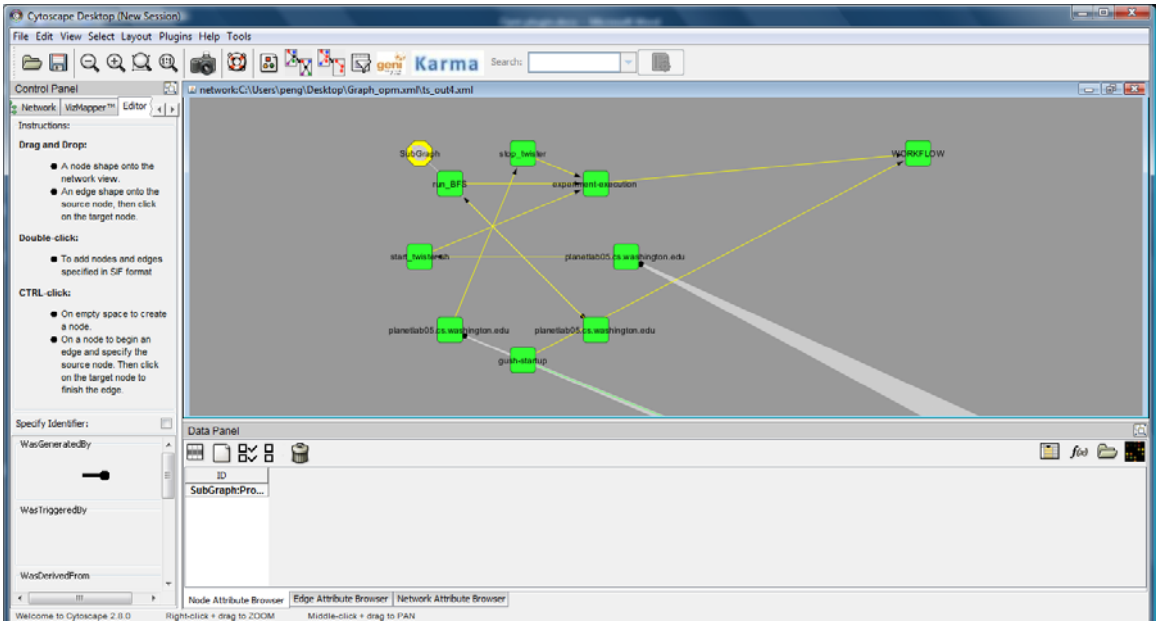

Figure 6.8.1.5 Cluster nodes

5) Repeat step 4 until you got an abstract graph showing only the smaller set of nodes you wish to visualize.

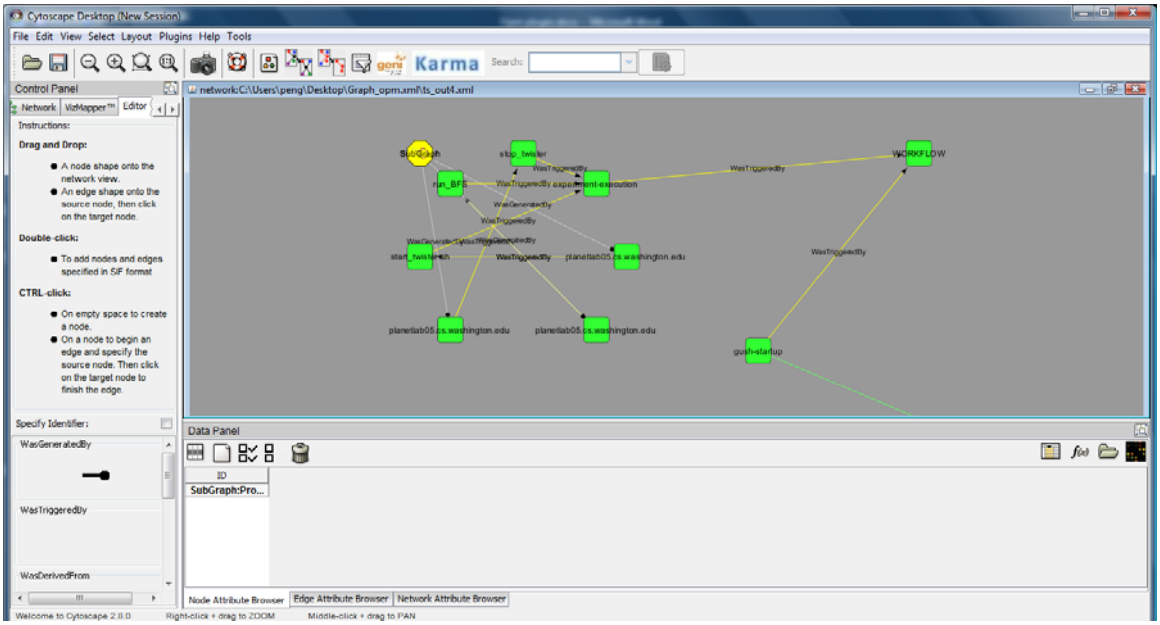

Figure 6.8.1.6 Create abstract graph

6.8.2 Navigate between graph and sub-graphs

Once you have created an abstract view, you can navigate between that parent abstract view and the sub-graph for any collapsed node.

1) Double click on any abstract node will show the view of the collapsed sub-graph

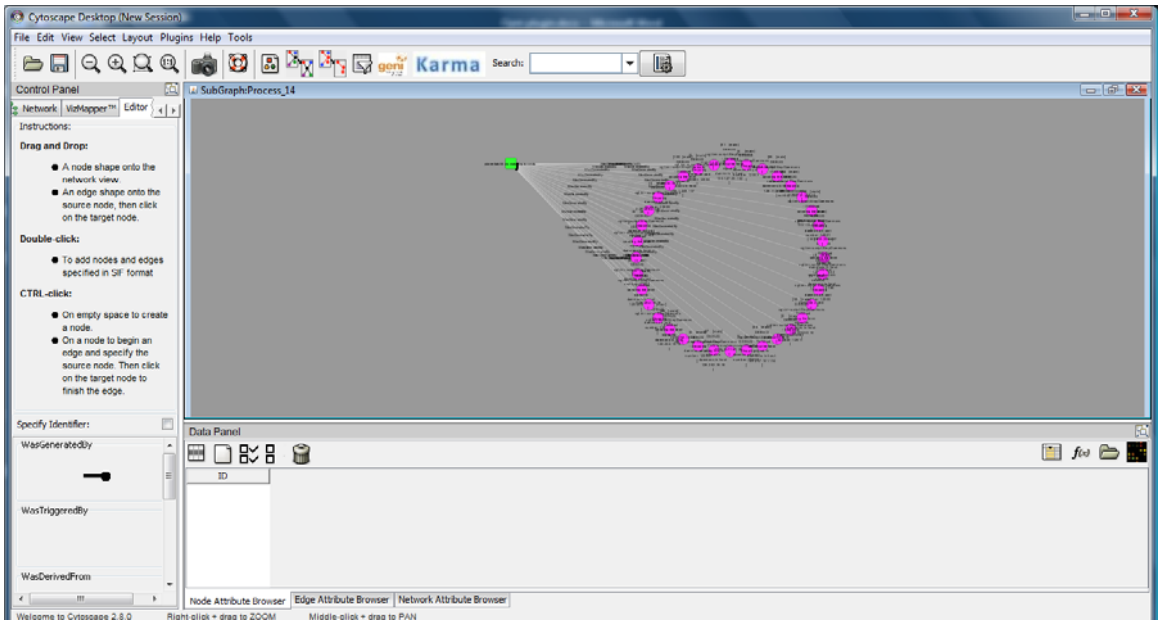
Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help Tools
Karma
Search:
Control Panel
Network VizMapper™ Editor
Instructions:
Drag and Drop:
A node shape onto the network view.
An edge shape onto the source node, then click on the target node.
Double-click:
To add nodes and edges specified in SIF format
CTRL-click:
On empty space to create a node.
On a node to begin an edge and specify the source node. Then click on the target node to finish the edge.
Specify Identifier:
WasGeneratedBy
WasTriggeredBy
WasDerivedFrom
SubGraph:Process_14
Data Panel
ID
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser

Figure 6.8.2.1 The sub-graph

2) You can go back to the parent graph by clicking on the item in the network panel on the left-hand side of the screen.

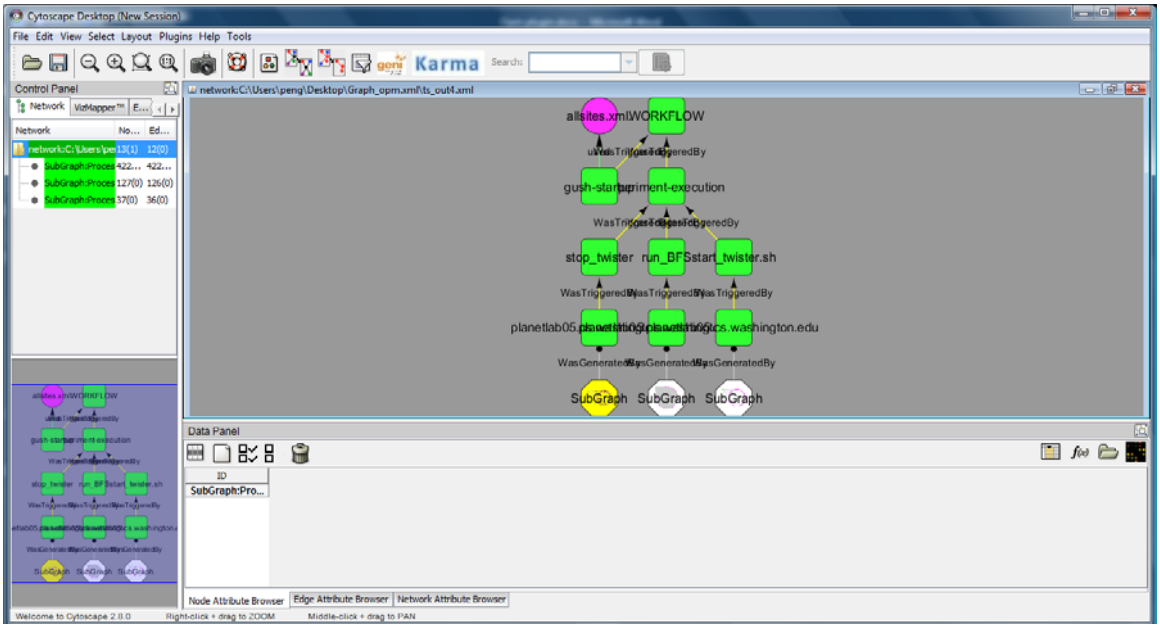

Figure 6.8.2.2 The parent graph

6.8.3 Compress Process/Artifact

Right clicking on any node in the graph, you will see menu options for “Compress Process” and “Compress Artifact”.

Clicking on “Compress Process”, the Cytoscape visualization plug-in will eliminate all of the “Process” nodes that link to two “artifact” nodes with the outgoing edge represents the relationship “used” and the incoming edge represents the relationship “wasGeneratedBy”. The process node will be replaced by a new edge between these two “artifact” nodes that represents the relationship “wasDerivedFrom”.

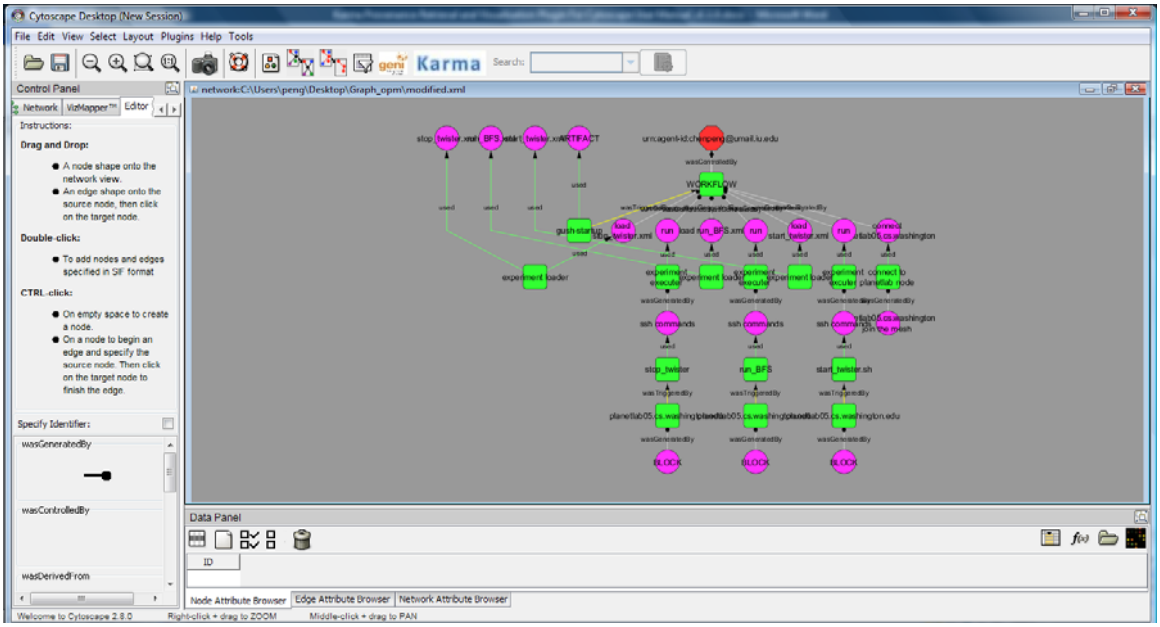

Figure 6.8.3.1 A provenance graph before applying “Compress Process”

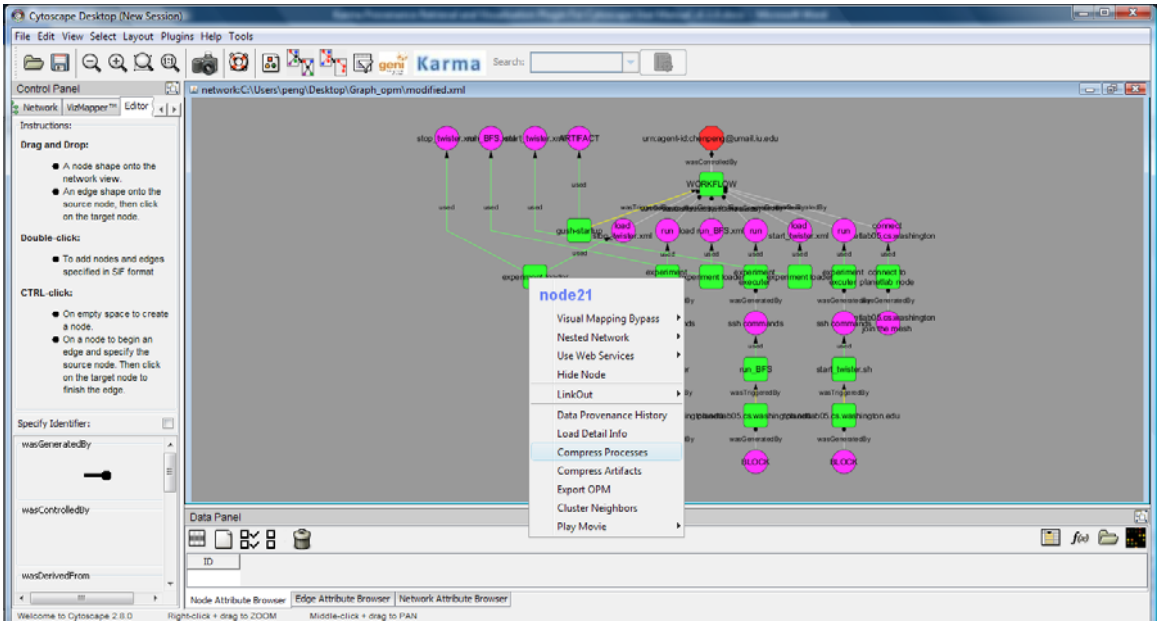

Figure 6.8.3.2 Applying “Compress Process”

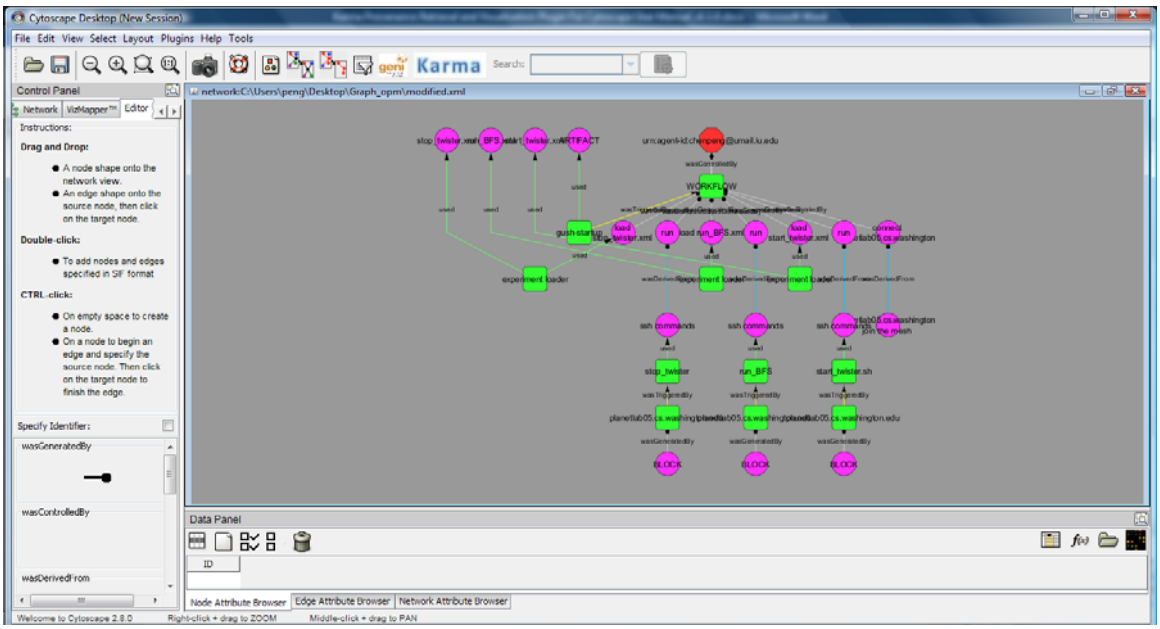
Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help Tools
Karma
Control Panel
Network VizMapper™ Editor
Instructions:
Drag and Drop:
A node shape onto the network view.
An edge shape onto the source node, then click on the target node.
Double-click:
To add nodes and edges specified in SIF format
CTRL-click:
On empty space to create a node.
On a node to begin an edge and specify the source node. Then click on the target node to finish the edge.
Specify Identifier:
wasGeneratedBy
wasControlledBy
wasDerivedFrom
Data Panel
ID
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.8.0
Right-click + drag to ZOOM
Middle-click + drag to PAN

Figure 6.8.3.3 The provenance graph after applying “Compress Process”

Clicking on “Compress Artifact”, the Cytoscape visualization plug-in will eliminate all of the “artifact” nodes that link two “Process” nodes where the outgoing edge represents the relationship “wasGeneratedBy” and the incoming edge represents the relationship “used”. The artifact will be replaced by a new edge between these two “Process” nodes that represents the relationship “wasTriggeredBy”.

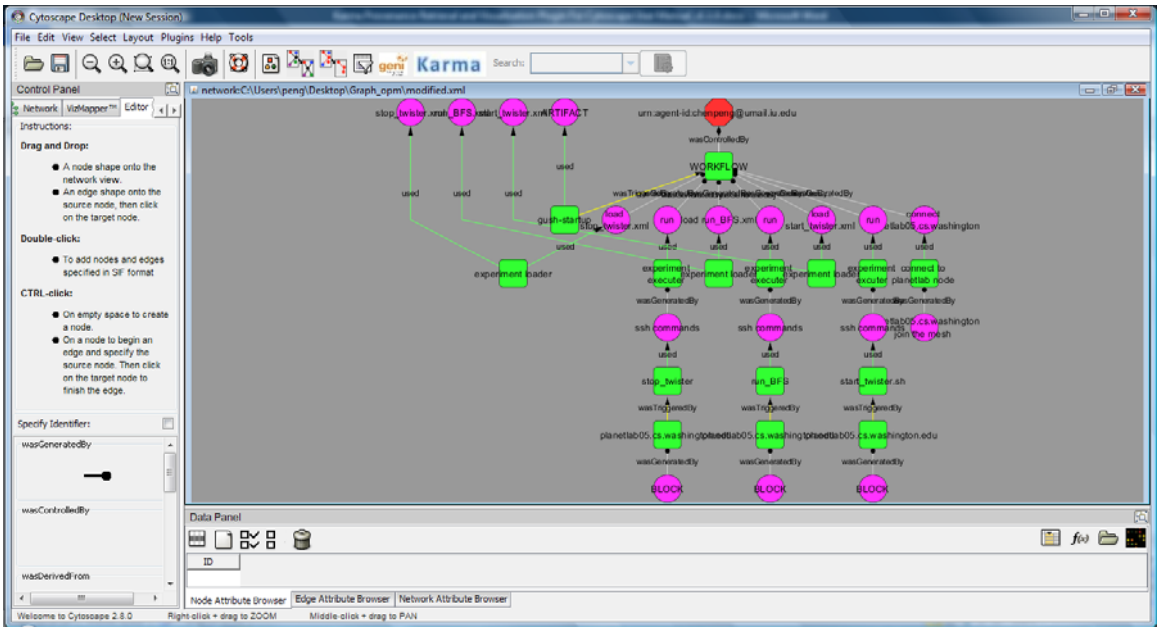
Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help Tools
Karma
Control Panel
Network VizMapper™ Editor
Instructions:
Drag and Drop:
A node shape onto the network view.
An edge shape onto the source node, then click on the target node.
Double-click:
To add nodes and edges specified in SIF format
CTRL-click:
On empty space to create a node.
On a node to begin an edge and specify the source node. Then click on the target node to finish the edge.
Specify Identifier:
wasGeneratedBy
wasControlledBy
wasDerivedFrom
networkC:\Users\peng\Desktop\Graph_opm\modified.xml
WORKFLOW
Data Panel
ID
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.8.0
Right-click + drag to ZOOM
Middle-click + drag to PAN

Figure 6.8.3.4 A provenance graph before applying “Compress Artifact”

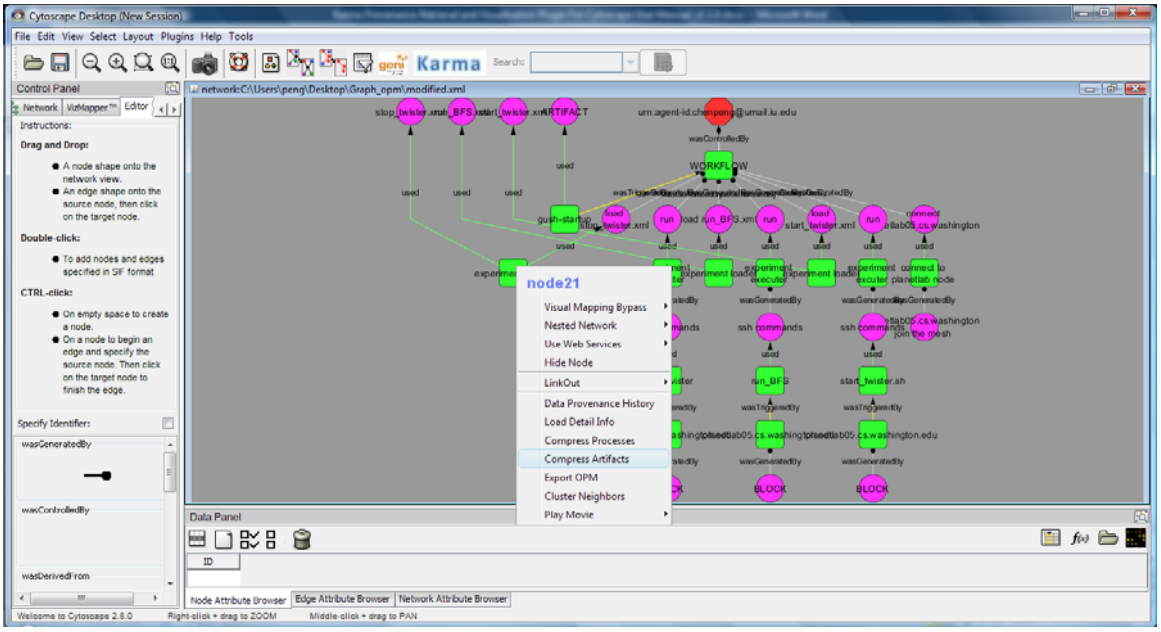

Figure 6.8.3.5 Applying “Compress Artifact”

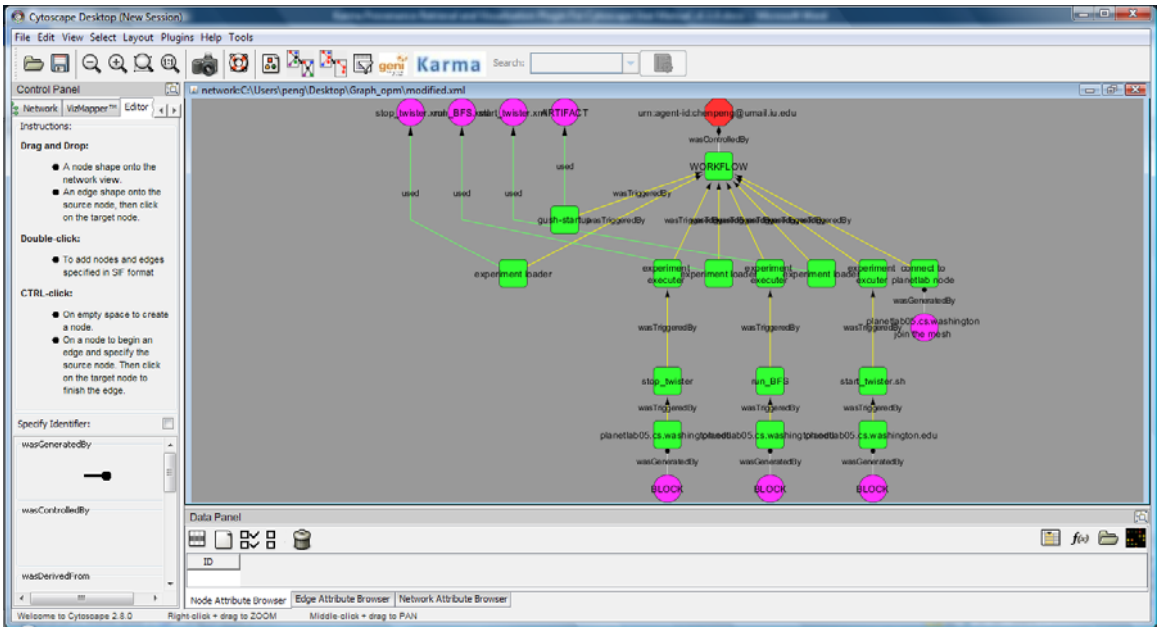

Figure 6.8.3.6 The provenance graph after applying “Compress Artifact”